Available online at www.sciencedirect.com
SCIENCE DIRECT®
Automatica 40 (2004) 2029–2041
automatica
www.elsevier.com/locate/automatica

Non-asymptotic quality assessment of generalised FIR models with periodic inputs☆

M.C. Campi[a], Su Ki Ooi[b], E. Weyer[b,*]

[a] *Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy*
[b] *CSSIP, Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia*

Received 11 August 2003; received in revised form 17 June 2004; accepted 21 June 2004
Available online 13 September 2004

Abstract

In any real-life identification problem, only a finite number of data points is available. On the other hand, almost all results in stochastic identification pertain to asymptotic properties, that is they tell us what happens when the number of data points tends to infinity. In this paper we consider the problem of assessing the quality of the estimates identified from a finite number of data points. We focus on least squares identification of generalised FIR models and develop a method to produce a bound on the uncertainty in the parameter estimate. The method is data driven and based on tests involving permuted data sets. Moreover, it does not require that the true system is in the model class.

Keywords: System identification; Model validation; Least squares; Finite samples properties; Confidence sets; Half samples

1. Introduction

The purpose of system identification is to construct a mathematical model of a dynamical system based on measurements. In order to give the user confidence in the obtained model, a quality tag should be delivered together with the model itself. Quality assessment is therefore an important issue and there has been a great deal of effort to derive methods for evaluating the model accuracy.

In this paper we consider the problem of assessing the quality of the parameter estimate obtained using least squares (LS) system identification methods with a finite number of data points. The mismatch between the true plant and the model consists of two components, bias error and variance error. In this work we focus on the variance error, which is due to the fact that the best model within the considered model class has not been found. If the best model within the model class is able to describe the true system, then the variance error is the only mismatch between the system and the model. However, if there are unmodelled dynamics, the variance error has to be combined with the bias error to obtain the total system-model mismatch. The cause of the bias error is that the model class considered is not rich enough to contain the 'true' plant. Works dealing with the problem of bounding the bias error include Goodwin, Gevers, and Ninness (1992), Hakvoort and Van den Hof (1997) and Ljung (2001).

In order for a method assessing the quality of the estimate to be useful, it must have the following properties. First, if we assume that the true system is in a given class, then the quality measure must be valid for all systems in that class. Furthermore, the quality measure must be computable based on the available a priori information about the true system and on the finite number of observed data points, and, finally, it must provide a rigorous result valid for the given number of data points.

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Soderstrom.
* Corresponding author. Tel.: +613-83449726; fax: +613-83446678.
E-mail addresses: campi@ing.unibs.it (M.C. Campi), skoo@ee.mu.oz.au (Su Ki Ooi), e.weyer@ee.mu.oz.au (E. Weyer).

0005-1098/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2004.06.020

The asymptotic convergence properties of the estimate are well understood, see e.g. (Ljung, 1999) or (Söderström & Stoica, 1988). Under natural conditions, if the true system belongs to the model class, the estimate $\hat{\theta}_N$ converges to θ_0 (the true parameter) with probability 1, and $\sqrt{N}(\hat{\theta}_N - \theta_0)$ is asymptotically normally distributed with zero mean and a certain variance P_{θ_0}. This result allows us to attach a quality tag to the estimate, and it is useful for gaining insight into the properties of system identification methods. However, with a finite number of data points the result holds only approximately and there are instances where application of the asymptotic theory leads to unreliable conclusions in the finite sample case, see e.g. (Bittanti, Campi, & Garatti, 2002) and (Garatti, Campi, & Bittanti, 2004).

Recently, the authors of this paper have studied finite sample properties of system identification methods in Weyer (2000), Campi and Weyer (2002) and Weyer and Campi (2002). The results obtained in those papers are essentially data-independent, in the sense that the bounds can be computed before any data are collected. This leads to results which are worst case with respect to the prior information and, consequently, the corresponding bounds may be conservative for the particular system at hand. In this paper, the properties of the estimate are studied after the data are collected, i.e. we obtain data-dependent results, leading to tighter evaluations.

We concentrate on the setting where the plant is identified through generalised FIR models with periodic inputs. This setting is of interest since periodic inputs are often used in applications when the input signal can be freely selected for identification purposes. Our set-up allows for the presence of unmodelled dynamics.

Finite sample properties have also been studied in different settings. Welsh and Goodwin (2002) have investigated the bias and variance of transfer function estimates obtained from indirect closed loop identification assuming a periodic reference signal and Gaussian noise. In the set membership and worst case identification setting, see e.g. (Wahlberg & Ljung, 1992; Giarre', Kacewicz, & Milanese, 1997; Chen & Gu, 2000; Milanese & Taragna, 2002), identification algorithms are constructed to deliver models which are in agreement with the observed data, and finite sample results are therefore delivered by the nature of the setting.

The paper is organised as follows. In Section 2, the idea of judging the quality of the estimate using half sample estimates is introduced and motivated. The model structure and the assumptions are precisely stated in Section 3, while Section 4 delivers the main result. A simulation example illustrating the developed method is presented in Section 5 followed by concluding remarks.

2. Quality assessment using half sample

In order to obtain a result of the type sought after, we propose to assess the quality of the LS estimate using half sample estimates. Assuming a linear regression predictor model $\hat{y}_{t,\theta} = \phi_t^{\mathrm{T}}\theta$, the LS estimate is given by

$$\hat{\theta}_N = (\Phi\Phi^{\mathrm{T}})^{-1}(\Phi Y),$$

where Y contains the outputs, $Y = [y_1, \ldots, y_N]^{\mathrm{T}}$ and Φ contains the regressors, $\Phi = [\phi_1, \ldots, \phi_N]$. Two half sample estimates

$$\theta' = (\Phi_1\Phi_1^{\mathrm{T}})^{-1}(\Phi_1 Y_1) \quad \text{and} \quad \theta'' = (\Phi_2\Phi_2^{\mathrm{T}})^{-1}(\Phi_2 Y_2)$$

can be computed using the first and second half of the data set, where (assuming N even)

$$Y_1 = [y_1, \ldots, y_{N/2}]^{\mathrm{T}}, \quad \Phi_1 = [\phi_1, \ldots, \phi_{N/2}],$$
$$Y_2 = [y_{(N/2)+1}, \ldots, y_N]^{\mathrm{T}}, \quad \Phi_2 = [\phi_{(N/2)+1}, \ldots, \phi_N].$$

Considering different ways of partitioning the data set in two halves, we can partition the original data set into two sets (Y_1, Φ_1) and (Y_2, Φ_2) with $N/2$ elements each in $\binom{N}{N/2}$ different ways. For example

$$Y_1 = [y_{N-1}, y_N, y_3, y_4, \ldots, y_{N/2}]^{\mathrm{T}},$$
$$\Phi_1 = [\phi_{N-1}, \phi_N, \phi_3, \phi_4, \ldots, \phi_{N/2}],$$
$$Y_2 = [y_{(N/2)+1}, \ldots, y_{N-2}, y_1, y_2]^{\mathrm{T}},$$
$$\Phi_2 = [\phi_{(N/2)+1}, \ldots, \phi_{N-2}, \phi_1, \phi_2]$$

is another possible partitioning. The idea is to judge the quality of the estimate $\hat{\theta}_N$ by the difference between these half sample estimates. The use of half sample estimates dates back a long time in the statistical literature, see e.g. (McCarthy, 1969) and (Hartigan, 1969).

If all values of the difference $\theta' - \theta''$ are within a small region around zero, we expect that $\hat{\theta}_N$ is a good estimate since there is little variation in the half sample estimates. On the other hand, we have a low confidence in $\hat{\theta}_N$ if the values of $\theta' - \theta''$ are widely spread. Intuitively, we take this as an indication that the variability due to noise, unmodelled dynamics, etc. have not been sufficiently averaged out and hence we do not place much confidence in the estimate.

A precise statement of the mathematical setting for the quality assessment method is postponed to Section 3. In this section, we illustrate some significant aspects and touch upon some issues of conceptual importance at a more intuitive level by examples.

2.1. Some preliminary examples

The first example illustrates that there are situations where any data-independent assessment (i.e. an assessment which can be computed before the data have been collected) of the model quality is impossible, and yet model dependent evaluations are possible. Thus, using data-dependent tests (of which half sample estimates methods are one example) is fundamental to derive sensible results.

Example 1. Consider the system

$$y_t = \theta_0 u_t + w_t, \tag{1}$$

where $u_t = 1$ for all t, and w_t is white Gaussian noise with zero mean and unknown variance σ^2. No upper bound on σ^2 is available. The predictor model used is $\hat{y}_{t,\theta} = \theta u_t$ and a set of input–output data, $D_N := \{(y_t, u_t)\}_{t=1}^N$ is collected from (1), where N is the number of data points. The LS estimate is $\hat{\theta}_N = (1/N)\sum_{t=1}^N y_t$. In this case $\hat{\theta}_N - \theta_0$ is a Gaussian random variable with zero mean and variance σ^2/N.

Suppose now that we want to make a statement of quality for the estimate that is data-independent, i.e. it can be evaluated before data have been collected and it holds for all possible true systems (i.e. for all possible values of σ^2). Precisely, the statement we are after is of the form

$$Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\} \geqslant 1 - \delta, \tag{2}$$

where ε (accuracy) and δ (confidence) are numbers that do not depend on the data. Since as already observed $\hat{\theta}_N - \theta_0$ is Gaussian $G(0, \sigma^2/N)$, for any given $\varepsilon > 0$ (even very large), $\sup_{\sigma^2} \delta(\sigma^2) = 1$, so that the only statement valid for all possible data generating system is

$$Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\} \geqslant 0,$$

which is evidently a void statement.

However, it is a well-known fact in statistics that a meaningful data-dependent quality statement can be made by resorting to the Student t-distribution. This illustrates the importance of using data not only in forming estimates but also in assessing their quality. Unfortunately, resorting to the Student t-distribution is possible in the simple context of this toy-example, whereas developing rigorous data-dependent results is far from simple in general.

It is well known (see e.g. (Richmond, 1964)) that

$$\frac{\hat{\theta}_N - \theta_0}{\sqrt{\frac{1}{N(N-1)}\sum_{t=1}^N (y_t - \hat{\theta}_N)^2}}$$

has a Student t-distribution with $N-1$ degrees of freedom (in the statistical literature this is called a 'pivotal' variable because its distribution does not depend on the unknown elements in the problem). Thus, given $\gamma > 0$, using the Student t-distribution table one can determine a δ such that

$$Pr\left\{\left(\frac{\hat{\theta}_N - \theta_0}{\sqrt{\frac{1}{N(N-1)}\sum_{t=1}^N (y_t - \hat{\theta}_N)^2}}\right)^2 \leqslant \gamma\right\} \geqslant 1 - \delta, \tag{3}$$

where the important fact is that (3) holds no matter what the data generating system is. In (3), γ and δ are data-independent. A data-dependent statement with the structure as in (2), viz.

$$Pr\left\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon(D_N)\right\} \geqslant 1 - \delta \tag{4}$$

can be readily derived from (3) with

$$\varepsilon(D_N) = \gamma \frac{1}{N(N-1)} \sum_{t=1}^N (y_t - \hat{\theta}_N)^2$$

(4) is the desired accuracy evaluation result: one selects a γ and computes the corresponding data-dependent accuracy parameter $\varepsilon(D_N)$. The Student t-distribution table is then used to find the associated confidence parameter δ.

The next example shows some preliminary facts regarding the parameter estimate quality assessment using half sample estimates.

Example 2. Consider the same situation as in Example 1. Let the number of data points N be even and split the index set $\{1, \ldots, N\}$ into two halves A_1 and A_2 with $N/2$ elements each. The two half sample estimates are given by

$$\theta' = \frac{1}{N/2} \sum_{t \in A_1} y_t \quad \text{and} \quad \theta'' = \frac{1}{N/2} \sum_{t \in A_2} y_t.$$

The difference $\theta' - \theta''$ is a zero mean Gaussian with variance $4\sigma^2/N$ and hence

$$Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\} = Pr\{(\theta' - \theta'')^2 \leqslant 4\varepsilon\}. \tag{5}$$

We can therefore evaluate the quality of the estimate in terms of the variation in the half sample estimates. Notice that the equality in (5) is valid for all σ^2 so it holds uniformly with respect to the data generating system.

In Example 2, $Pr\{(\theta' - \theta'')^2 \leqslant 4\varepsilon\}$ is of course unknown since it depends on the true system. Nevertheless, by partitioning the data into two subsets in, say, M different ways, we obtain M different pairs of half sample estimates θ_i' and θ_i'', $i = 1, \ldots, M$, and we can estimate $p = Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\}$ by

$$\hat{p}(D_N) = \frac{1}{M} \sum_{i=1}^M \mathbf{1}((\theta_i' - \theta_i'')^2 \leqslant 4\varepsilon), \tag{6}$$

where $\mathbf{1}$ is the indicator function, i.e. the estimate $\hat{p}(D_N)$ is the empirical frequency of the event $(\theta' - \theta'')^2 \leqslant 4\varepsilon$. $\hat{p}(D_N)$ is an unbiased estimator for p. However, the claim $Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\} = \hat{p}(D_N)$ is itself stochastic since the estimate $\hat{p}(D_N)$ is data-dependent, and the claim does not make sense at a conceptual level. In order to make sense, the claim has to be qualified with a second probability giving us the probability that the claim itself is true. The next example illustrates this matter, and it is a special case of our main result Theorem 4.7.

Example 3. Consider the same situation as in Examples 1 and 2. Assume that we have split the data set into two subsets in M different ways such that $\theta_i' - \theta_i''$, $i = 1, \ldots, M$, are iid, independent and identically distributed (see Section 4

for how this can be achieved). Let $\hat{p}(D_N)$ be as in (6). Then the statement

$$Pr\{(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon\} \geqslant \hat{p}(D_N) - \rho$$

holds true with probability at least $1 - \mathrm{e}^{-2M\rho^2}$. Here ρ is a margin of error on the estimated probability. In words, the claim is that the probability of the squared estimation error to be less than or equal to ε is greater than or equal to $\hat{p}(D_N) - \rho$. However, this claim is itself probabilistic since $\hat{p}(D_N) - \rho$ is a random variable. The second probability tells us that the claim itself is true with probability not less than $1 - \mathrm{e}^{-2M\rho^2}$. This probability goes to 1 rapidly as the last term is exponential in M, so that it can be neglected in many practical situations. Yet, it is important to observe that disregarding this probability leads to an unsound mathematical statement. Moreover, making this probability close to 1 comes at the cost of decreasing the probability that $(\hat{\theta}_N - \theta_0)^2 \leqslant \varepsilon$. This tradeoff is represented by the "tuning" parameter ρ.

In the next section we introduce the approach of assessing the model quality using half sample estimates in a general setting.

3. The identification setting

3.1. Model class and input signal

We consider models with predictors of the type

$$\hat{y}_{t,\theta} = \theta_1 g_1(q^{-1}, u_t) + \cdots + \theta_n g_n(q^{-1}, u_t).$$

This predictor can be written in linear regression form $\hat{y}_{t,\theta} = \phi_t^{\mathrm{T}}\theta$ with

$$\theta = [\theta_1, \ldots, \theta_n]^{\mathrm{T}},$$
$$\phi_t = [g_1(q^{-1}, u_t), \ldots, g_n(q^{-1}, u_t)]^{\mathrm{T}}.$$

Here θ is the parameter vector to be estimated, q^{-1} is the backward shift operator (i.e. $q^{-1}u_t = u_{t-1}$), and $g_k(q^{-1}, u_t)$, $k = 1, \ldots, n$, is a short form for $g_k(u_t, u_{t-1}, \ldots)$, which are linear or non-linear functions of the past inputs.

Example. A popular choice of $g_k(q^{-1}, u_t)$ is $L_k(q^{-1}, \alpha)u_t$, where $L_k(q^{-1}, \alpha) = \frac{\sqrt{1-\alpha^2}}{q-\alpha}\left(\frac{1-\alpha q}{q-\alpha}\right)^{k-1}$ are the Laguerre polynomials and α is a parameter to be chosen by the user.

Next we introduce the assumptions on the input signal.

A1. The input signal is deterministic and periodic with period L. Moreover, $g_k(u_t, u_{t-1}, \ldots)$, $k = 1, \ldots, n$, are well-defined i.e., when the actual input is substituted, the g_k's return a finite value.

The assumption that the g_k's are well-defined is a mild assumption that relates to the stability of the g_k operators. When computing $g_k(u_t, u_{t-1}, \ldots)$, we have in principle to substitute the periodic input u up to time t starting back from $-\infty$, implicitly assuming that the periodic input has been applied since time $-\infty$. In practice, the periodic input u is applied for long enough so that the tail behavior in the g_k functions is negligible.

Suppose that the true system output has been observed for N periods, i.e. N periods are used in identification. To simplify notation let

$$\Phi = [\phi_1, \ldots, \phi_{NL}] = [\Phi_1, \ldots, \Phi_1],$$
$$\Phi_1 = [\phi_1, \ldots, \phi_L] = [\phi_{(i-1)L+1}, \ldots, \phi_{iL}],$$
$$i = 1, \ldots, N,$$

where the last equality is due to the periodicity of the input.

As a final assumption on the input signal we assume that:

A2. The input u_t and the functions $g_k(q^{-1}, u_t)$ are chosen such that $\Phi\Phi^{\mathrm{T}}$ is non-singular, i.e. the LS estimate is unique.

3.2. True system

We assume that the true system can be written as

$$y_t = h(q^{-1}, u_t) + w_t,$$

which in vector form becomes

$$Y = \bar{Y} + W,$$

where

$$Y = [y_1, \ldots, y_{NL}]^{\mathrm{T}} = [Y_1^{\mathrm{T}}, \ldots, Y_N^{\mathrm{T}}]^{\mathrm{T}},$$
$$\bar{Y} = [h(q^{-1}, u_1), \ldots, h(q^{-1}, u_{NL})]^{\mathrm{T}} = [\bar{Y}_1^{\mathrm{T}}, \ldots, \bar{Y}_N^{\mathrm{T}}]^{\mathrm{T}},$$
$$W = [w_1, \ldots, w_{NL}]^{\mathrm{T}} = [W_1^{\mathrm{T}}, \ldots, W_N^{\mathrm{T}}]^{\mathrm{T}},$$
$$Y_i = [y_{(i-1)L+1}, \ldots, y_{iL}]^{\mathrm{T}}, \quad i = 1, \ldots, N,$$
$$\bar{Y}_i = [h(q^{-1}, u_{(i-1)L+1}), \ldots, h(q^{-1}, u_{iL})]^{\mathrm{T}}, \ i = 1, \ldots, N,$$
$$W_i = [w_{(i-1)L+1}, \ldots, w_{iL}]^{\mathrm{T}}, \quad i = 1, \ldots, N.$$

Here $h(q^{-1}, u_t) = h(u_t, u_{t-1}, \ldots)$ is a causal operator of past input signals. We assume that

A3. $h(q^{-1}, u_t) = h(u_t, u_{t-1}, \ldots)$ is well-defined, i.e., when the actual input is substituted, h returns finite values.

A4. $\Phi_1 W_i$, $i = 1, \ldots, N$, are iid and symmetrically distributed around zero.

Assumptions A3 and A4 deserve some words of explanation.

Note first that writing $y_t = h(u_t, u_{t-1}, \ldots) + w_t$ subsumes that the system has been initialized in the remote past with the periodic input u, which in practice means that the transients have died out. Requiring that $h(u_t, u_{t-1}, \ldots)$ is well-defined is necessary since if e.g. the true system is unstable the output generated by h can as well escape to infinity. More subtle is the observation that assuming that the true system output is, up to noise, given by $h(q^{-1}, u_t) = h(u_t, u_{t-1}, \ldots)$ corresponds implicitly to assume that, after transients have died out, the system outputs a periodic signal when fed

by a periodic input signal u (note in fact that expression $h(u_t, u_{t-1}, \ldots)$ returns the same value at time t and time $t+L$). This assumption is very mild and is satisfied e.g. by all asymptotically stable linear time invariant systems. Many non-linear systems exhibit this behavior as well, provided that they meet certain requirements of stability. A trivial example is given by an asymptotically stable linear system with static input and output non-linearities. We also note that, due to the periodicity of h, $\bar{Y}_1 = \bar{Y}_2 = \cdots = \bar{Y}_N$ in the definition of $\bar{Y}$.

As for the noise, Assumption A4 is certainly satisfied if w_t is a sequence of iid random variables symmetrically distributed around zero. Assumption A4 also allows for correlated noise as long as $\Phi_1 W_i$ is iid over blocks of data. Moreover, in situations where correlation in the noise prevents A4 from being rigorously satisfied, this assumption is still expected to hold approximately due to the averaging effect of the product $\Phi_1 W_i$. Indeed, the elements of vector $\Phi_1 W_i$ are weighted sums of all the w_t variables over a period. Assuming that the time dependence in w_t is short as compared to the period L, even in two adjacent periods these weighted sums will contain many terms that are almost independent, so that the sums themselves will be almost independent. This is important for the—at least approximate—applicability of the results in this paper to practical situations. We also remark that our results can be extended to dependent, mixing $\Phi_1 W_i$'s. However, the mixing case is significantly more involved, and a full presentation of results in the mixing setting would detract from the principal ideas, and is therefore not justified. See Weyer (2000) or Weyer and Campi (1999) for results using mixing conditions.

Importantly, our setting does not assume that the true system is contained in the model class.

3.3. LS estimate

The LS estimate is given by

$$\hat{\theta}_{NL} = (\Phi\Phi^{\mathrm{T}})^{-1}(\Phi Y). \tag{7}$$

As the number of data points tends to infinity, $\hat{\theta}_{NL}$ converges to

$$\theta^* = (\Phi\Phi^{\mathrm{T}})^{-1}(\Phi E Y) = (\Phi\Phi^{\mathrm{T}})^{-1}(\Phi\bar{Y}),$$

where E is the expectation operator.

The following lemmas will be used in subsequent derivations.

Lemma 3.1. *Y can be written as $Y = \Phi^{\mathrm{T}}\theta^* + \bar{W}$ where $\bar{W}$ has the property that $\Phi\bar{W} = \Phi W$.*

Proof. $\Phi\bar{W} = \Phi(Y - \Phi^{\mathrm{T}}\theta^*) = \Phi(\bar{Y} + W - \Phi^{\mathrm{T}}\theta^*) = \Phi W$. □

The difference between $\hat{\theta}_{NL}$ and θ^* can be expressed in terms of W and Φ as shown in the next lemma.

Lemma 3.2.

$$(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) = W^{\mathrm{T}}\Phi^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})^{-1}\Phi W.$$

Proof. $\hat{\theta}_{NL} - \theta^* = (\Phi\Phi^{\mathrm{T}})^{-1}(\Phi Y) - \theta^* = (\Phi\Phi^{\mathrm{T}})^{-1}\Phi\bar{W} = (\Phi\Phi^{\mathrm{T}})^{-1}\Phi W$, where the last equality follows from Lemma 3.1 Hence, $(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) = W^{\mathrm{T}}\Phi^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})^{-1}\Phi W$. □

4. An algorithm for model quality assessment using half sample estimates

The data blocks $\{(Y_i, \Phi_1)\}_{i=1}^N$ are split into two subsets $\{(Y_i, \Phi_1)\}_{i\in A_1}$ and$\{(Y_i, \Phi_1)\}_{i\in A_2}$, where $A_1 = \{i_1, \ldots, i_{N/2}\}$ and $A_2 = \{j_1, \ldots, j_{N/2}\}$ are two disjoint index sets containing $N/2$ elements each (N is assumed even). In other words, each data block corresponds to one period, and we have split the data block set into two subsets of equal size, each containing $N/2$ data blocks. Let θ' and θ'' denote the half sample LS estimates computed on each of the two subsets, i.e.

$$\theta' = 2(\Phi\Phi^{\mathrm{T}})^{-1}\sum_{i\in A_1}\Phi_1 Y_i \quad \text{and}$$
$$\theta'' = 2(\Phi\Phi^{\mathrm{T}})^{-1}\sum_{i\in A_2}\Phi_1 Y_i.$$

Next we introduce some notation. Let $\beta = [\bar{\beta}_1, \ldots, \bar{\beta}_N]^{\mathrm{T}}$ be an N-vector with

$$\bar{\beta}_i = \begin{cases} 1 & i \in A_1, \\ -1 & i \in A_2 \end{cases}$$

and let

$$H_\beta = \begin{bmatrix} \bar{\beta}_1 I_L & 0 & \ldots & 0 \\ 0 & \bar{\beta}_2 I_L & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \bar{\beta}_N I_L \end{bmatrix},$$

where I_L is the $L \times L$ identity matrix.

The two next lemmas show that the quality of the estimate $\hat{\theta}_{NL}$ can be related to the difference between the two half sample estimates.

Lemma 4.1.

$$(\theta' - \theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta' - \theta'') = 4W^{\mathrm{T}}H_\beta\Phi^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})^{-1}\Phi H_\beta W.$$

Proof.

$$\begin{aligned}\theta' - \theta'' &= 2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\sum_{i\in A_1}\Phi_1 Y_i - \sum_{i\in A_2}\Phi_1 Y_i\right)\\ &= 2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\Phi_1\sum_{i=1}^{N}\bar{\beta}_i Y_i\right)\\ &= 2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\Phi_1\sum_{i=1}^{N}\bar{\beta}_i(\bar{Y}_i + W_i)\right)\\ &= 2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\Phi_1\sum_{i=1}^{N}\bar{\beta}_i W_i\right)\\ &= 2(\Phi\Phi^{\mathrm{T}})^{-1}\Phi H_\beta W, \end{aligned} \tag{8}$$

where the second last equality follows since $\sum_{i=1}^{N}\bar{\beta}_i = 0$ and $\bar{Y}_1 = \bar{Y}_2 = \cdots = \bar{Y}_N$, from which the lemma easily follows. □

Lemma 4.2.

$$\begin{aligned}&Pr\{(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon\}\\ &\quad = Pr\{(\theta' - \theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta' - \theta'') \leqslant 4\varepsilon\}.\end{aligned}$$

Proof. From Assumption A4, $\Phi W = \sum_{i=1}^{N}\Phi_1 W_i$ and $\Phi H_\beta W = \sum_{i=1}^{N}\Phi_1\bar{\beta}_i W_i$ have the same distribution. The lemma then follows from Lemmas 3.2 and 4.1. □

In view of Lemmas 4.1 and 4.2, an algorithm for estimating

$$p = Pr\{(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon\}$$

is to partition the data blocks into two subsets, each containing $N/2$ blocks, in a number of different ways and then estimate p as the frequency of the event $\{(\theta' - \theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta' - \theta'') \leqslant 4\varepsilon\}$ where the half sample estimates are computed from the partitioned data sets.

Theorem 4.3 below formalises this idea. Before the theorem is stated, some additional notation is required. Let $\beta_j = [\bar{\beta}_{j,1}, \ldots, \bar{\beta}_{j,N}]^{\mathrm{T}}$, $j = 1, \ldots, M$, be M vectors with half of the elements equal to 1 and the other half equal to -1 and let $\mathbf{1}((\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j - \theta''_j) \leqslant 4\varepsilon)$ be the indicator function of the event $(\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j - \theta''_j) \leqslant 4\varepsilon$ where

$$\theta'_j - \theta''_j = 2(\Phi\Phi^{\mathrm{T}})^{-1}\sum_{i=1}^{N}\Phi_1\bar{\beta}_{j,i}Y_i$$

is the difference between the half sample estimates obtained when the data is partitioned into two sets according to whether $\bar{\beta}_{j,i}$ is $+1$ or -1. (That is all data blocks (Φ_1, Y_i) with i such that $\bar{\beta}_{j,i} = 1$ is in one set and the blocks (Φ_1, Y_i) corresponding to $\bar{\beta}_{j,i} = -1$ is in the other data set.)

Theorem 4.3. *Given a model class and a true system as in Sections* 3.1 *and* 3.2, *let the LS estimate be given by* (7), *and let* $\beta_1, \ldots, \beta_M$ *be N-vectors with* $N/2$ *entries equal to* 1 *and* $N/2$ *entries equal to* -1. *Then,*

$$\begin{aligned}\hat{p}(D_{NL}) = \frac{1}{M}\sum_{j=1}^{M}&\mathbf{1}((\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})\\ &\times (\theta'_j - \theta''_j) \leqslant 4\varepsilon)\end{aligned} \tag{9}$$

is an unbiased estimator for p.

Proof.

$$\begin{aligned}E\,\hat{p}(D_{NL}) &= \frac{1}{M}\sum_{j=1}^{M}E\mathbf{1}((\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})\\ &\quad\times (\theta'_j - \theta''_j) \leqslant 4\varepsilon)\\ &= \frac{1}{M}\sum_{j=1}^{M}Pr\{(\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})\\ &\quad\times (\theta'_j - \theta''_j) \leqslant 4\varepsilon\}\\ &= \frac{1}{M}\sum_{j=1}^{M}Pr\{(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})\\ &\quad\times (\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon\} = p,\end{aligned}$$

where Lemma 4.2 has been used in the second last step. □

Theorem 4.3 delivers a way of assessing the model quality: estimate the probability p using the expression (9) for $\hat{p}(D_{NL})$. As $\hat{p}(D_{NL})$ is an unbiased estimate of p, this returns a non-conservative evaluation of the probability that $(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon$.

However, one needs to be careful with the interpretation of the estimate $\hat{p}(D_{NL})$. The statement

$$Pr\{(\hat{\theta}_{NL} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon\} = \hat{p}(D_{NL}) \tag{10}$$

makes no sense at a conceptual level since $\hat{p}(D_{NL})$ is data-dependent and hence stochastic. Statement (10) needs to be qualified with another level of probability, giving us the probability that the stochastic quality claim (10) is true. What is sought is a bound on the probability that $p \geqslant \hat{p}(D_{NL}) - \rho$ where ρ is a margin. We have the following result.

Theorem 4.4. *If* $\mathbf{1}((\theta'_j - \theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j - \theta''_j) \leqslant 4\varepsilon)$, $j = 1, \ldots, M$, *are independent of each other, then*

$$Pr\{p \geqslant \hat{p}(D_{NL}) - \rho\} \geqslant 1 - \mathrm{e}^{-2M\rho^2}.$$

Proof. The proof is based on Hoeffding's inequality. This inequality states that, if v_j, $j = 1, 2, \ldots, M$, are independent random variables taking value in $[0, 1]$, then

$$Pr\{E[S_M] \geqslant S_M - \rho\} \geqslant 1 - \mathrm{e}^{-2M\rho^2},$$

where $S_M = \frac{1}{M}\sum_{j=1}^{M}v_j$ (see (Vidyasagar, 1997) for details).

Thus, if we identify v_j with $\mathbf{1}\left((\theta'_j-\theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j-\theta''_j)\leqslant 4\varepsilon\right)$, by the independence property of these variables the result follows.

We further show that independence of the $\mathbf{1}((\theta'_j-\theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j-\theta''_j)\leqslant 4\varepsilon)$ variables can be secured by choosing vectors β_i's that are mutually orthogonal, under some additional conditions. Admittedly, these conditions are restrictive and others may find more general settings in which independence holds.

We commence with two preliminary lemmas and provide the second level of probability in the subsequent Theorem 4.7.

Lemma 4.5. *Let $N=2^l$, for some integer l. Then, there exist $N-1$ mutually orthogonal vectors of size N whose elements are 1 and -1 with an equal number of each.*

Proof. The proof is by induction. For $l=1$, the vector $a_{1,1}=[1\ \ -1]^{\mathrm{T}}$ satisfies the claim of the Lemma.

Assume that the claim is true for $N=2^l$ for some $l\geqslant 1$ and name the vectors $a_{l,1}, a_{l,2}, \ldots, a_{l,2^l-1}$. Then, for $N=2^{l+1}$ the following $N-1$ vectors are mutually orthogonal

$$
\begin{aligned}
a_{l+1,1} &= [1,\ldots,1,\ \ -1,\ldots,-1]^{\mathrm{T}},\\
a_{l+1,2} &= [a_{l,1}^{\mathrm{T}},\ \ a_{l,1}^{\mathrm{T}}]^{\mathrm{T}},\\
a_{l+1,3} &= [a_{l,2}^{\mathrm{T}},\ \ a_{l,2}^{\mathrm{T}}]^{\mathrm{T}},\\
&\vdots\\
a_{l+1,N-2^l} &= [a_{l,2^l-1}^{\mathrm{T}},\ \ a_{l,2^l-1}^{\mathrm{T}}]^{\mathrm{T}},\\
a_{l+1,N-2^l+1} &= [-a_{l,1}^{\mathrm{T}},\ \ a_{l,1}^{\mathrm{T}}]^{\mathrm{T}},\\
&\vdots\\
a_{l+1,N-1} &= [-a_{l,2^l-1}^{\mathrm{T}},\ \ a_{l,2^l-1}^{\mathrm{T}}]^{\mathrm{T}}. \qquad \square
\end{aligned}
$$

We also note that there cannot be more than $N-1$ orthogonal vectors, since N orthogonal vectors would form a basis for R^N, which is impossible since vectors with an equal number of 1 and -1 can only span a subspace of vectors whose entries add up to zero.

Lemma 4.6. *Let $\beta_1,\ldots,\beta_M$ be mutually orthogonal. Strengthen Assumption* A4 *to the following Assumption* A4′:

A4′. *w_t is a sequence of iid zero mean Gaussian random variables.*

Then,

$$Pr\{p\geqslant \hat{p}(D_{NL})-\rho\}\geqslant 1-\mathrm{e}^{-2M\rho^2}.$$

Proof. Based on Theorem 4.4, we have to show that $\mathbf{1}((\theta'_j-\theta''_j)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_j-\theta''_j)\leqslant 4\varepsilon)$ and $\mathbf{1}((\theta'_k-\theta''_k)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'_k-\theta''_k)\leqslant 4\varepsilon)$ are independent for $j\neq k$.

The indicator function $\mathbf{1}\left((\theta'-\theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'-\theta'')\leqslant 4\varepsilon\right)$ is a measurable function of $\theta'-\theta''$, and hence the two indicator functions above are independent if $\theta'_j-\theta''_j$ is independent of $\theta'_k-\theta''_k$. Under Assumption A4′, $\theta'_j-\theta''_j$ and $\theta'_k-\theta''_k$ are zero mean Gaussian random vectors, hence they are independent if they are uncorrelated. We therefore have to prove that $E(\theta'_j-\theta''_j)(\theta'_k-\theta''_k)^{\mathrm{T}}=0$. Now, from (8) we have

$$
\begin{aligned}
&E(\theta'_j-\theta''_j)(\theta'_k-\theta''_k)^{\mathrm{T}}\\
&\quad= E4(\Phi\Phi^{\mathrm{T}})^{-1}(\Phi H_{\beta_j}W)(W^{\mathrm{T}}H_{\beta_k}\Phi^{\mathrm{T}})(\Phi\Phi^{\mathrm{T}})^{-1}\\
&\quad= 4(\Phi\Phi^{\mathrm{T}})^{-1}[\Phi H_{\beta_j}(EWW^{\mathrm{T}})H_{\beta_k}\Phi^{\mathrm{T}}](\Phi\Phi^{\mathrm{T}})^{-1}\\
&\quad= 4\sigma^2(\Phi\Phi^{\mathrm{T}})^{-1}(\Phi H_{\beta_j}H_{\beta_k}\Phi^{\mathrm{T}})(\Phi\Phi^{\mathrm{T}})^{-1}.
\end{aligned}
$$

Observing that $\Phi H_{\beta_j}=[\bar{\beta}_{j,1}\Phi_1,\ldots,\bar{\beta}_{j,N}\Phi_1]$, we obtain $\Phi H_{\beta_j}H_{\beta_k}\Phi^{\mathrm{T}}=\sum_{i=1}^N\bar{\beta}_{j,i}\bar{\beta}_{k,i}\Phi_1\Phi_1^{\mathrm{T}}=\Phi_1\Phi_1^{\mathrm{T}}\sum_{i=1}^N\bar{\beta}_{j,i}\bar{\beta}_{k,i}=\Phi_1\Phi_1^{\mathrm{T}}\beta_j^{\mathrm{T}}\beta_k=0$, since β_j and β_k are orthogonal, so completing the proof. $\square$

In view of Lemma 4.6 we have the following theorem where the existence of the β_j vectors follows from Lemma 4.5.

Theorem 4.7. *Given a model class and a true system as in Sections* 3.1 *and* 3.2, *let the number of observed periods of data be $N=2^l$, for some integer l, and let the LS estimate be given by* (7). *Consider the estimator $\hat{p}(D_{NL})$ defined by* (9) *and assume that $\beta_1,\ldots,\beta_M$ are mutually orthogonal where $M=N-1$, and that* A4′ *is satisfied. Then, the statement*

$$Pr\{(\hat{\theta}_{NL}-\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta^*)\leqslant\varepsilon\}\geqslant\hat{p}(D_{NL})-\rho$$

holds true with probability no smaller than $1-\mathrm{e}^{-2(N-1)\rho^2}$.

The above theorem involves two levels of probability. Firstly, it claims that the probability that $\hat{\theta}_{NL}$ and θ^* are less than a certain distance apart is larger than or equal to $\hat{p}(D_{NL})-\rho$. This claim is itself stochastic since $\hat{p}(D_{NL})$ is a random variable. The second level of probability tells us that the quality claim is true with probability at least $1-\mathrm{e}^{-2(N-1)\rho^2}$. For example, with $\rho=0.1$, and $N=128$ the statement holds true with probability no less than 0.92113. The probability rapidly increases with N, if $\rho=0.1$ and $N=512$ the probability that the statement holds true is already at least 0.99996. This is the effect of the exponential function. Decreasing the value of ρ leads instead to a rapid rise in the required number of periods. This behavior is in the nature of things and is a well-known fact in the related field of statistical learning. The reason why the external probability $1-\mathrm{e}^{-2(N-1)\rho^2}$ goes to 1 exponentially with N is that we have wisely selected the β_i

vectors to be orthogonal, so that the different terms entering the estimate $\hat{p}(D_{NL})$ are independent (see proof of Lemma 4.6).

Theorem 4.7 enables one to compute the quality of the parameter estimate based on the observed data and it provides a rigorous result valid for a finite number of data points. Since $\hat{p}(D_{NL})$ is an unbiased estimator of p, the only place any conservativeness is introduced is in the second probability through the margin ρ. Notice that the Gaussian assumption A4′ is only used when bounding the second probability. $\hat{p}(D_{NL})$ is still an unbiased estimator under the weaker assumption A4.

Theorem 4.3 provides a result for LS estimation with periodic inputs which holds under general technical conditions. In Theorem 4.7, the technical conditions have been strengthened by a significant degree by assuming iid, Gaussian noise. Yet, unmodelled dynamics is allowed.

Remark 1. Under the conditions of Theorem 4.7 a confidence ellipsoid with a data dependent ε can be obtained similarly to Example 1 in Section 2.1 and therefore the conditions of Theorem 4.7 are indeed restrictive. In fact,

$$\frac{(\hat{\theta}_{NL}-\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta^*)}{n\hat{\sigma}^2}, \tag{11}$$

where $\hat{\sigma}^2=\frac{1}{NL}\sum_{t=1}^{\frac{NL}{2}}t\left(y_t-y_{t+\frac{NL}{2}}\right)^2$ is a pivotal variable which is Fisher distributed $F\left(n,\frac{NL}{2}\right)$. If one further restricts generality by assuming that the true system belongs to the model class, one obtains the standard result ((Ljung, 1999), Appendix II) that

$$\frac{(\hat{\theta}_{NL}-\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta^*)}{n\hat{\sigma}^2_{NL}} \tag{12}$$

has a Fisher $F(n,NL-n)$ distribution, where $\hat{\sigma}^2_{NL}=\frac{1}{NL-n}\sum_{t=1}^{NL}(y_t-\phi_t^{\mathrm{T}}\hat{\theta}_{NL})^2$.

4.1. Relationship to other methods and techniques

Though the method developed in this paper presents its own specific characteristics, it shares common aspects with methodologies and techniques used in other related fields, and a brief discussion of some of these techniques now follows.

4.1.1. Bootstrap

Notice that we can express y_t as $y_t=\phi_t^{\mathrm{T}}\hat{\theta}_{NL}+\varepsilon_{t,\hat{\theta}_{NL}}$ where $\varepsilon_{t,\hat{\theta}_{NL}}=y_t-\hat{y}_{t,\hat{\theta}_{NL}}=y_t-\phi_t^{\mathrm{T}}\hat{\theta}_{NL}$ is the prediction error. Introducing the notation $\Xi_{i,\theta}=[\varepsilon_{(i-1)L+1,\theta},\ldots,\varepsilon_{iL,\theta}]^{\mathrm{T}}$, $i=1,\ldots,N$, we have that

$$\begin{aligned}\theta'-\theta''&=2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\sum_{i\in A_1}\Phi_1(\Phi_1^{\mathrm{T}}\hat{\theta}_{NL}+\Xi_{i,\hat{\theta}_{NL}})\right.\\&\qquad\left.-\sum_{i\in A_2}\Phi_1(\Phi_1^{\mathrm{T}}\hat{\theta}_{NL}+\Xi_{i,\hat{\theta}_{NL}})\right)\\&=2(\Phi\Phi^{\mathrm{T}})^{-1}\Phi_1\left(\sum_{i\in A_1}\Xi_{i,\hat{\theta}_{NL}}-\sum_{i\in A_2}\Xi_{i,\hat{\theta}_{NL}}\right)\\&=2(\Phi\Phi^{\mathrm{T}})^{-1}\left(\Phi_1\sum_{i=1}^{N}\bar{\beta}_i\Xi_{i,\hat{\theta}_{NL}}\right),\end{aligned}$$

where the $\bar{\beta}_i$'s are defined as at the beginning of Section 4. Thus, the difference between half sample estimates is nothing but a suitably weighted average of the prediction errors.

Basic implementations of bootstrap would be based on a random resampling from $\varepsilon_{t,\hat{\theta}_{NL}}$. The technique of partitioning blocks of output data in two subsets in a number of different ways bears some resemblances with the resampling technique used in bootstrap. However, in our approach the partitioning is done in a systematic and deterministic fashion, and unlike bootstrap there is no random sampling from an empirical distribution. One way to view our proposed method is that we have replaced the original problem of estimating $Pr\{(\hat{\theta}_{NL}-\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta^*)\leqslant\varepsilon\}$ with the problem of estimating $Pr\{(\theta'-\theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'-\theta'')\leqslant 4\varepsilon\}$ where the latter problem is "easier" since we have $N-1$ iid realisations of $(\theta'-\theta'')^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\theta'-\theta'')$ at hand.

Performing a systematic and deterministic partition of blocks of data has an important advantage over random resampling in that we are not forced to model the data generation mechanism in detail (so that $\varepsilon_{t,\hat{\theta}_{NL}}$ is white) as it is the case in basic implementations of bootstrap. Our approach allows for unmodelled dynamics as well since the unmodelled dynamics is cancelled out by the way the $\bar{\beta}_i$'s coefficients are chosen. For details on bootstrap in a system identification setting see e.g. (Tjärnström & Ljung, 2002).

4.1.2. Subsampling

The approach is also connected with subsampling methods (Politis, Romano, & Wolf, 1999) where the quality in the estimate is assessed by comparing the estimate with estimates computed on subsets of the total data set. This can be seen by noting that $\theta'-\theta''=2(\hat{\theta}_{NL}-\theta'')$. Hartigan (1969) has used subsamples to compute exact confidence intervals for a scalar parameter.

4.1.3. Rademacher sequences

The technique of changing signs of data blocks has similarities with the use of Rademacher sequences in learning theory, see e.g. (Koltchinskii & Panchenko, 2000; Koltchinskii, 2001) or (Mendelson, 2002). A Rademacher sequence

is an iid sequence $\{r_i\}$, where each r_i take on the value 1 or -1 with probability 1/2 each. In function learning one tries to learn an unknown function f based on n observations of f at iid extracted points $x_1, \ldots, x_n$. Typical results involving Rademacher sequences in function learning are of the form

$$Pr\left\{\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - Ef\right| \geqslant \varepsilon\right\} \leqslant 4Pr\left\{\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} r_i f(x_i)\right| \geqslant \frac{n\varepsilon}{4}\right\},$$

where the probability on the left is with respect to $x_1, \ldots, x_n$ and the probability on the right with respect to $x_1, \ldots, x_n$ and the Rademacher sequence. Empirical evaluations of the right-hand side are often referred to as Rademacher bootstrap.

Our approach of changing sign can be viewed as making use of "deterministic Rademacher sequences".

4.1.4. Multitaper power spectrum estimation

The approach of selecting matrices H_{β_j} such that the half sample estimates are independent of each other has similarities with the way windowing functions are selected in the multitaper approach to spectral estimation. The idea is to window the time domain data by orthogonal window functions such that the obtained spectral estimates (periodograms) are independent of each other and then average the periodograms to reduce the variance. For more details see e.g. (Manolakis, Ingle, & Kogon, 2000).

4.1.5. Permutation tests

The proposed method for model quality assessment has some features in common with permutation tests in statistical hypotheses testing. For more details on permutation tests see e.g. (Lehmann, 1986; Van der Vaart, 1998) or (Good, 2000). Although there are similarities between permutation tests and our proposed method, the purpose of permutation tests is quite different from the objective here: in our proposed method the permutations are used as a tool via the computation of half sample estimates for assessing the quality of the obtained model, while in permutation tests they are used for testing statistical hypotheses.

4.2. Discussion

From Section 4.1.1, it is clear that $\theta' - \theta''$ can be expressed in terms of the prediction errors. A considerable amount of work has been done in assessing model quality using the prediction errors. In particular, Ljung and Guo (1997) have studied what a typical model validation test based on the prediction errors implies in terms of the model error, expressed in the frequency domain. Moreover, a procedure for estimating probabilistic uncertainty regions, which involves the explicit calculation of the bias and variance errors of a linear regression estimate, has been developed in (Hakvoort & Van den Hof, 1997) and (Goodwin et al., 1992). In the latter paper, stochastic embedding is used to produce an estimate of the mean square error between the true and estimated nominal transfer function.

In the present paper, only the variance error has been considered, but we could have extended the method to take into account the bias along the lines of the above cited papers. However, this would have introduced conservativeness, and our results are most useful in situations where we have a priori information that the bias error is small in the frequency range of interest, i.e. the variance error is the one that contributes the most to the total error in that frequency range. This is illustrated in the simulation example in the next section.

The extension to non-periodic input signals appears more difficult. Some of the difficulties in moving to the non-periodic case along a rigorous route are to find matrices H_β such that the statistical properties of ΦW are the same as those of $\Phi H_\beta Y$ and to find an unbiased estimator of p. This is due to the fact that in the non-periodic case the unmodelled dynamics cannot be averaged out. Furthermore, the orthogonality condition is not preserved and hence Hoeffding's inequality cannot be used to bound the outer probability. However, in the special case where we have only one parameter, we have been able to generalise the results to non-periodic input signals along a rigorous route, see (Ooi, Weyer, & Campi, 2003).

5. Simulations

Consider the following system:

$$y_t = a_1^0 y_{t-1} + a_2^0 y_{t-2} + b_1^0 u_{t-1} + b_2^0 u_{t-2} + w_t \tag{13}$$

with parameter $\theta_0 = [a_1^0, a_2^0, b_1^0, b_2^0]^{\mathrm{T}} = [1.4, -0.45, 0.07, 0.04]^{\mathrm{T}}$ (the poles are at 0.5 and 0.9) and w_t is Gaussian white noise with zero mean and variance $\sigma^2 = 0.34$. The model class is the following second order Laguerre model class

$$\hat{y}_{t,\theta} = \theta_1 L_1(q^{-1}, \alpha)u_t + \theta_2 L_2(q^{-1}, \alpha)u_t,$$

where $\theta = [\theta_1, \theta_2]^{\mathrm{T}}$ is a parameter to be estimated and $L_k(q^{-1}, \alpha) = \frac{\sqrt{1-\alpha^2}}{q-\alpha}(\frac{1-\alpha q}{q-\alpha})^{k-1}$, $k = 1, 2$, $\alpha = 0.85$, is an a priori available estimate of the system dominant pole 0.9.

A multi-sine input signal, $u(t)$ with five different frequencies in addition to a DC component is generated, i.e.

$$u_t = \sum_{k=1}^{5} \frac{1}{10}\sin(\omega_k t + \varrho_k) + 0.2,$$

where ω_k, for $k = 1, \ldots, 5$, are $\frac{2\pi}{125}$, $\frac{4\pi}{125}$, $\frac{8\pi}{125}$, $\frac{12\pi}{125}$ and $\frac{16\pi}{125}$, and the sinusoids have Schroeder phases, $\varrho_k = -\frac{k(k-1)\pi}{5}$ in order to keep their crest factor small, see (Ljung, 1999).

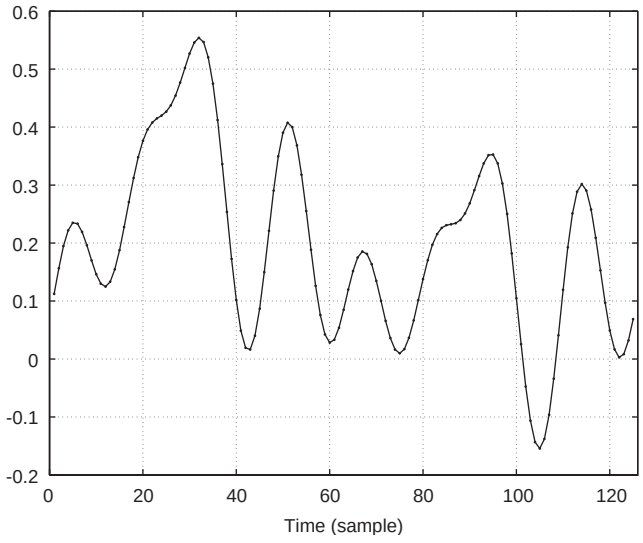

Fig. 1. Plot of one period of the multi sine input signal.

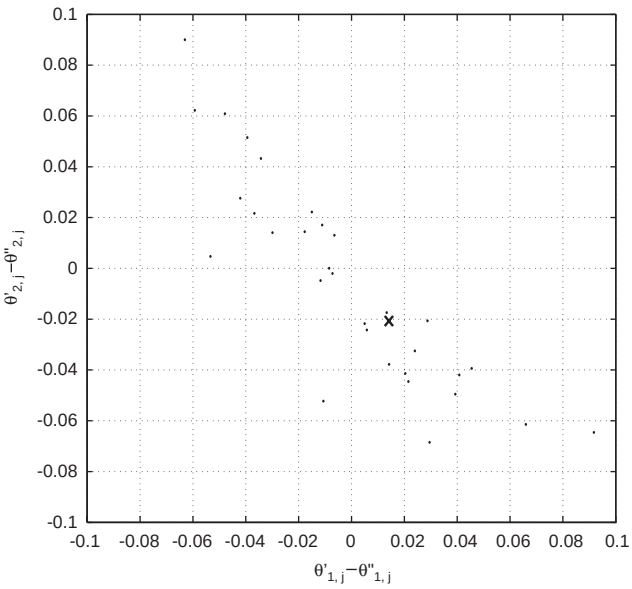

Fig. 2. Scatter plot of $\theta'_j - \theta''_j$, $j = 1, \ldots, 31$. '×' denotes $\hat{\theta}_{4000} - \theta^*$.

Fig. 1 shows the plot of one period (i.e. $L = 125$) of the input signal. The number of periods is $N=32$ and $NL=4000$ data points are generated according to system (13). For this example, $\theta^* = [0.4223,\ 0.2032]^{\mathrm{T}}$.

First, $\hat{\theta}_{4000}$ is computed. Then, ε is fixed to 2 and an estimate $\hat{p}(D_{4000})$ of $p = Pr\{(\hat{\theta}_{4000} - \theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{4000} - \theta^*) \leqslant \varepsilon\}$ is obtained from the simulated data with mutually orthogonal β_j's. The whole process is then repeated another 199 times so as to compare the probability $1-\mathrm{e}^{-2(N-1)\rho^2}$ of Theorem 4.7 with the corresponding empirical probability that $p \geqslant \hat{p}(D_{4000}) - \rho$.

5.1. Results

The scatter plot of $\theta'_j - \theta''_j$, $j=1, \ldots, 31$, for a simulation together with the true value of $\hat{\theta}_{4000} - \theta^*$ denoted by a cross is displayed in Fig. 2. In this simulation we obtained $\hat{\theta}_{4000} = [0.4365, 0.1825]^{\mathrm{T}}$.

From this scatter plot, it is observed that the half sample estimates form a region around zero, and that $\hat{\theta}_{4000} - \theta^*$ is within this region, illustrating that assessing the quality of the estimate using half sample estimates is a feasible approach.

The elements of the matrix $\Phi\Phi^{\mathrm{T}} = \left[\begin{smallmatrix} 2566 & 2161 \\ 2161 & 2566 \end{smallmatrix}\right]$ are large compared to $\varepsilon = 2$, so the estimation error in each component of the parameter vector is small. The magnitude of the elements in $\Phi\Phi^{\mathrm{T}}$ increases "linearly" with N and, when the value of ε is chosen, the magnitude of the entries in $\Phi\Phi^{\mathrm{T}}$ should be taken into account.

Fig. 3 shows the plot of the ellipsoid: $(\hat{\theta}_{4000} - \theta)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{4000} - \theta) \leqslant \varepsilon$ with $\varepsilon = 2$.

The empirical distribution of $\hat{p}(D_{4000})$ obtained over the total of 200 repetitions of the simulation is plotted in Fig. 4 together with the true value of $p = Pr\{(\hat{\theta}_{NL} -$

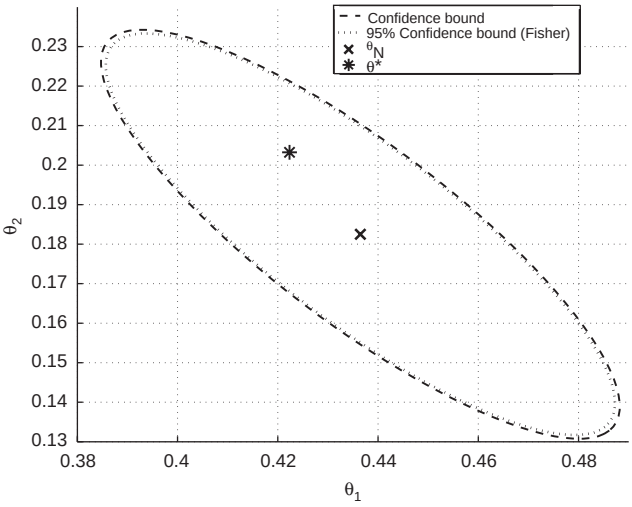

Fig. 3. Plot of $(\hat{\theta}_{NL} - \theta)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta) \leqslant 2$ and 95% confidence bound obtained using the Fisher distribution, '×' denotes $\hat{\theta}_{4000}$ and '*' denotes θ^*.

$\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL} - \theta^*) \leqslant \varepsilon\} = 0.9472$ denoted by a cross. As expected, the true value of p falls in a central position among the $\hat{p}(D_{4000})$'s since $\hat{p}(D_{4000})$ is unbiased and hence a non-conservative estimator of p.

By using the Fisher distribution, as detailed in Remark 1, we obtained the 95% confidence ellipsoid $\frac{(\hat{\theta}_{NL}-\theta)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta)}{n\hat{\sigma}^2} \leqslant 3.0002$ as shown in Fig. 3, where $\hat{\sigma}^2$ is estimated as

$$\hat{\sigma}^2 = \frac{1}{NL}\sum_{t=1}^{\frac{NL}{2}}(y_t - y_{t+\frac{NL}{2}})^2 = 0.3212.$$

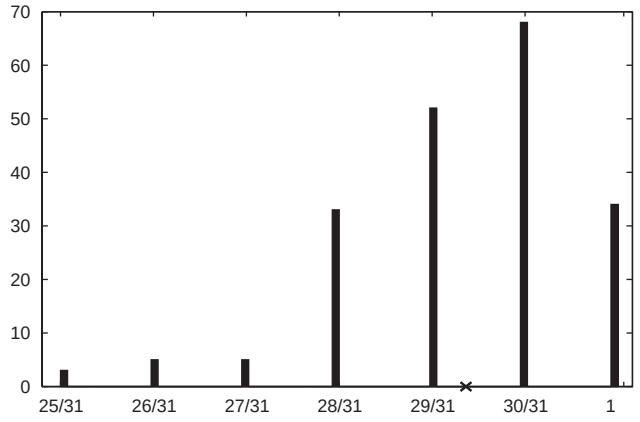

Fig. 4. Empirical distribution of $\hat{p}(D_{4000})$.

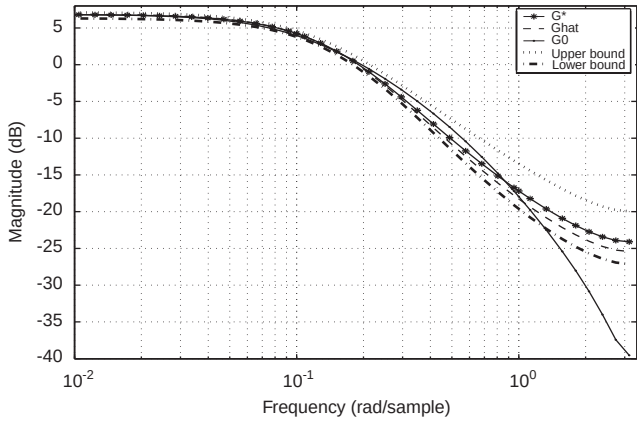

Fig. 5. Frequency domain plot ($G0$: True system, *Ghat*: Estimated model and G^*: "Best model").

Table 1
μ_H and μ_s for different values of ρ

N	L	ε	ρ	μ_H	μ_s
32	125	2	$\frac{1}{31}$	0.0625	0.83
			0.1	0.4621	1
			0.3	0.9962	1
			0.5	0.9999	1

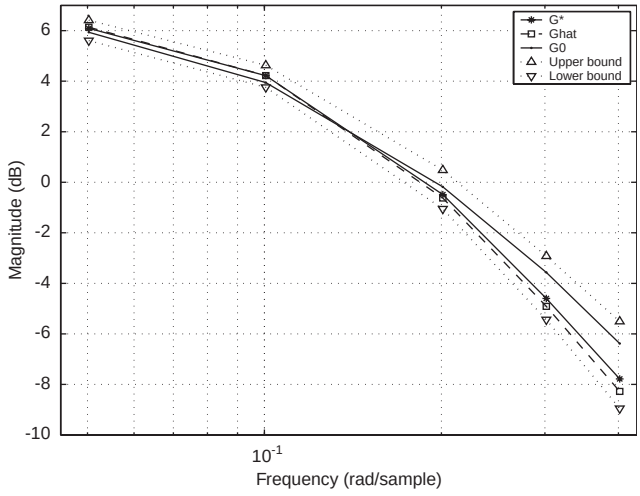

Fig. 6. Frequency domain plot: Zoomed in version.

We observe that the two confidence ellipsoids are very similar, further illustrating that $\hat{p}(D_{4000})$ is an unbiased estimate.

Next, let $\mu_H = 1 - e^{-2(N-1)\rho^2}$ denote the probability we guarantee the quality claim with ("H" stands for "Hoeffding", since this bound is computed from Hoeffding's inequality). μ_H is computed for different values of ρ, and compared with the empirical frequency that $p \geqslant \hat{p}(D_{4000}) - \rho$ (μ_s in Table 1, where "s" stands for "sampling"). From Table 1, when $\rho = \frac{1}{31}$ we place little confidence in the obtained statement about the quality of the estimate even though it actually held true for 166 out of the 200 simulations. This shows that in this example the second level of probability is conservative. Notice however that this is the only place where conservativeness is introduced as $\hat{p}$ is an unbiased estimator. μ_H can be increased by increasing ρ. The price to pay is a more conservative bound for $Pr\{(\hat{\theta}_{4000} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{4000} - \theta^*) \leqslant \varepsilon\}$. We see that there is a natural trade off between the additional margin ρ in the bound on $Pr\{(\hat{\theta}_{4000} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{4000} - \theta^*) \leqslant \varepsilon\}$ and the confidence in the claim about the quality of the estimate.

The uncertainty in the parameter estimate was transferred to the frequency domain in order to obtain a bound on the transfer function. The frequency domain plots of the true system, the estimated model, the uncertainty bounds and the best model are given in Fig. 5. Fig. 6 shows the zoomed in version of the frequency domain plot with markers on the frequencies of the input signal.

In the simulation of Figs. 5 and 6, $\hat{p}(D_{4000}) = 0.9677$, and when $\rho = 0.1$, the result in Table 1 tells us that 86.77% of the time θ^* will be located within the ellipsoid, or equivalently the best model transfer function will be situated within the uncertainty region in Fig. 5, and this claim is true with probability larger than 0.4621. Obviously there is a fair bit of conservativeness in the outer probability.

From Fig. 5 it is clear that the uncertainty region is quite small in the low frequency range. The uncertainty region not only covers the variance error but it also includes the transfer function of the true system since the bias error is small. From Fig. 6 we observe that the estimated model is very close to the best model at the frequencies of the input signals. As these are the frequencies where the input energy is located, the bias error is also quite small at these frequencies. From Fig. 5 we can see that the bias error increases in the high frequency region.

The above results show that the proposed method for model quality assessment delivers useful frequency domain bounds, particularly in the frequency range where the bias error is small.

6. Conclusion

In this paper we have presented new results on model quality assessment of system identification models. We have considered LS estimation of generalised FIR models with periodic inputs. Importantly, we have not assumed that the true system belongs to the model class. The probability $p=Pr\{(\hat{\theta}_{NL}-\theta^*)^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}})(\hat{\theta}_{NL}-\theta^*)\leqslant\varepsilon\}$ is estimated using an unbiased and hence non-conservative estimator based on permutations of the data set. As the estimate of p is stochastic, a second probability is needed in order to assert the probability with which the stochastic quality claim is true. This second probability is obtained using Hoeffding's inequality. The bound on the quality of the parameter estimate as stated in Theorem 4.7 provides a rigorous result valid for a finite number of data points.

These results are less conservative than previous finite sample results. Simulation results have shown that the proposed method works well and that we can transfer the uncertainty in the parameter estimates into an uncertainty region in the frequency domain.

Acknowledgements

This work is partly supported by the European Commission under the project HYBRIDGE IST-2001-32460, and by MIUR under the project "New methods for Identification and Adaptive Control for Industrial Systems".

This research has been supported by the Cooperative Research Centre for Sensor Signal and Information Processing under the Cooperative Research Centre scheme funded by The Commonwealth Government.

The authors are indebted to an anonymous reviewer for many useful comments that helped improve this paper.

References

Bittanti, S., Campi, M. C., & Garatti, S. (2002). New result on the asymptotic theory of system identification for the assessment of the quality of estimated models. *Proceedings of the 41st IEEE CDC, Las Vegas, Nevada, USA.* (pp. 1814–1819).

Campi, M. C., & Weyer, E. (2002). Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, *47*(8), 1329–1334.

Chen, J., & Gu, G. (2000). Control oriented system identification. An H_∞ approach. New York: Wiley.

Garatti,, S., Campi, M. C., & Bittanti, S. (2004). Assessing the quality of identified models through the asymptotic theory—When is the result reliable?. *Automatica*, *40*(8), 1319–1332.

Giarre', L., Kacewicz, B., & Milanese, M. (1997). Model quality evaluation in H_2 identification. *Automatica*, *33*, 1133–1139.

Good, P. (2000). Permutation tests: a practical guide to resampling methods for testing hypotheses. 2nd ed., Berlin: Springer.

Goodwin, G. C., Gevers, M., & Ninness, B. (1992). Quantifying the errors in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, *37*(7), 913–928.

Hakvoort, R. G., & Van den Hof, P. M. (1997). Identification of probabilistic system uncertainty regions by explicit evaluation of bias and variance errors. *IEEE Transactions on Automatic Control*, *42*(11), 1516–1528.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, *64*(328), 1303–1317.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions Information Theory*, *47*(6), 1902–1914.

Koltchinskii, V., & Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. in: Gine, Mason, & Wellner (Eds.), *High dimensional probability II*. (pp. 443–459). Basel: Birkhäuser.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. 2nd ed., New York: Wiley, Reprinted (1994) edition by Chapman and Hall.

Ljung, L. (1999). *System identification theory for the user*. 2nd ed., Englewood Cliffs: Prentice-Hall.

Ljung, L. (2001). Estimating linear time invariant models of non-linear time-varying systems. *European Journal of Control*, *7*(2–3), 203–219.

Ljung, L., & Guo, L. (1997). The role of model validation for assessing the size of the unmodeled dynamics. *IEEE Transactions on Automatic Control*, *42*(9), 1230–1239.

Manolakis, D. G., Ingle, V. K., & Kogon, S. M. (2000). *Statistical and adaptive signal processing*. New York: McGraw-Hill.

McCarthy, P. J. (1969). Pseudo-replication: half samples. *Review of The International Statistical Institute*, *37*, 239–263.

Mendelson, S. (2002). Rademacher averages and phase transition in Glivenko-Cantelli classes. *IEEE Transactions Information Theory*, *48*(1), 251–263.

Milanese, M., & Taragna, M. (2002). Optimality, approximation and complexity in set membership H_∞ identification. *IEEE Transactions on Automatic Control*, *47*(10), 1682–1690.

Ooi, S. K., Weyer, E., & Campi, M. C. (2003). Finite sample quality assessment of system identification models of irrigation channels. *Proceedings of the IEEE conference on control application, Istanbul, Turkey*.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Berlin: Springer.

Richmond, S. (1964). *Statistical analysis*. 2nd ed., New York: The Ronald Press Company.

Söderström, T., & Stoica, P. (1988). *System identification*. Englewood Cliffs: Prentice-Hall.

Tjärnström, F., & Ljung, L. (2002). Using the Bootstrap to estimate the variance in the case of undermodeling. *IEEE Transactions on Automatic Control*, *47*(2), 395–398.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Vidyasagar, M. (1997). *A theory of learning and generalization with applications to neural networks and control systems*. Berlin: Springer.

Wahlberg, B., & Ljung, L. (1992). Hard frequency-domain model error bounds from least-squares like identification techniques. *IEEE Transactions on Automatic Control*, *AC-37*, 900–912.

Welsh, J. S., & Goodwin, G. C. (2002). Finite sample properties of indirect nonparametric closed-loop identification. *IEEE Transactions on Automatic Control*, *47*(8), 1277–1292.

Weyer, E. (2000). Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, *36*, 1291–1299.

Weyer, E., & Campi, M. C. (1999). Finite sample properties of system identification methods. *Proceedings of the 38th IEEE CDC, Phoenix, Arizona, USA.* (pp. 510–515).

Weyer, E., & Campi, M. C. (2002). Non-asymptotic confidence ellipsoids for the least squares estimate. *Automatica*, *38*(9), 1539–1547.

Marco Claudio Campi is Professor of Automatic Control at the University of Brescia, Italy.

He was born in Tradate, Italy, on December 7, 1963. In 1988, he received the Doctor degree in electronic engineering from the Politecnico di Milano, Milano, Italy (his doctoral thesis was awarded the "Giorgio Quazza" prize as the best original thesis for year 1988). From 1988 to 1989, he was a Research Assistant at the Department of Electrical Engineering of the Politecnico di Milano. From 1989 to 1992, he worked as a Researcher at the Centro di Teoria dei Sistemi of the National Research Council (CNR) in Milano. Since 1992, he has been with the University of Brescia, Italy.

M.C. Campi is an Associate Editor of Automatica and Systems and Control Letters, and a past Associate Editor of the European Journal of Control. Serves as Chair of the Technical Committee IFAC on Stochastic Systems (SS) and is a member of the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP). Moreover, he is a Distinguished Lecturer under the IEEE Control System Society (CSS) Program. He has held visiting and teaching positions at many universities and institutions including the Australian National University, Canberra, Australia, the University of Illinois at Urbana-Champaign, USA, the Centre for Artificial Intelligence and Robotics, Bangalore, India, and the University of Melbourne, Australia.

The research interests of M.C. Campi include: adaptive and data-based control, system identification, robust convex optimization, robust control and estimation, and learning theory. His research activity has been conducted through years under many Italian and European projects. Currently, he is leader of the Brescia unit of the European IST project "Distributed control and stochastic analysis of hybrid systems supporting safety critical real-time systems design". The research activity of M.C. Campi is witnessed by more than 40 papers published in archival journals.

Su Ki Ooi was born in Penang, Malaysia, in 1976. He obtained the B.Eng. and Ph.D. degrees in Electrical and Electronic Engineering both from The University of Melbourne, Australia, in 1999 and 2004, respectively.

Since 2004 he has been a Research Fellow in the Department of Electrical and Electronic Engineering, The University of Melbourne, Australia. His current research interests are in the area of system identification and control. He is a member of the IEEE and the IEAust.

Erik Weyer received the Siv. Ing. degree in 1988 and the Ph.D. in 1993, both from the Norwegian Institute of Technology, Trondheim Norway. During his Ph.D. studies he spent two years at the Australian National University.

From 1994 to 1996 he was a Research Fellow at the University of Queensland, and since 1997 he has been with the Department of Electrical and Electronic Engineering, the University of Melbourne, where he is currently a Senior Research Fellow.

His research interests are in the area of system identification and control.